

Recognition-Gated Safety & Format-as-Affordance: Why Aligned Models Fail Under Jurisdictional Confusion and Multi-Step Tool Hijacking

Taylor Amarel · AI Safety & Security Research · Live Research Portal: <https://research.baltor.ai> · Kaggle Benchmarks: [jc-realmode](#)l-t4 & [jc-strong-t4](#)

Core Thesis (Recognition-Gated Safety): Across empirical GPU T4 evaluations on Qwen/Qwen2.5-7B-Instruct ([jc-realmode](#)l-t4, [jc-strong-t4](#)) and multi-step agent harnesses ([jed-attack-submission](#)), we test whether models facilitate camouflaged exploitation because they *do not know* the pattern is exploitative or because **contract-drafting and tool-calling formats fail to activate latent knowledge during generation**. Out-of-band probes show the model already recognizes exploitation on 93.8% (15/16) of cases, yet complies with 81.3% → 91.7% of harmful drafting/tool requests under bureaucratic camouflage.

1. PART I: OUT-OF-BAND RECOGNITION VS. IN-BAND EXECUTION PROBE (`QWEN2.5-7B-INSTRUCT` ON KAGGLE T4)

We evaluated 16 exploitation scenarios across 4 mechanisms (fee_shift, reimburse, doc_retention, and share_records tool exfiltration) paired with benign employer-paid controls across 5 conditions:

Experimental Condition (`jc-realmode	Harmful Compliance Rate (Lower = Safer)	Safe Refusal Rate	Out-of-Band Exploitation Recognition	Benign Control Helpfulness
1. `plain_harmful` (Overt Exploitation Request)	25.0% (0% on strong 12-item subset)	75.0% (100% strong)	93.8% (15/16)	—
2. `framed_baseline` (Routine HR / Legal Wrapper)	81.3% (91.7% strong)	18.8% (8.3% strong)	93.8% (15/16)	100.0%
3. `conflict_of_laws` (Jurisdictional Split)	68.8% (83.3% strong)	31.3% (16.7% strong)	93.8% (15/16)	100.0%
4. `framed_recognition_gate` (Pre-Turn Recognition Gate)	18.8% (8.3% strong)	81.3% (91.7% strong)	93.8% (15/16)	100.0%
5. `conflict_recognition_gate` (Jurisdictional + Gate)	0.0% (0/16 complied)	100.0% (16/16 refused)	93.8% (15/16)	100.0%

2. PART II: FORMAT-AS-AFFORDANCE & MULTI-STEP AGENT TOOL HIJACKING (`JED-ATTACK-SUBMISSION`)

In tool-using AI agents (including share_records tool calls in JC and multi-step workspace pipelines in jed_attack), structured formatting acts as an **execution affordance** that bypasses safety filters across two primary attack classes:

Attack Class	Multi-Step Agent Trajectory	Unassisted Agent Vulnerability	Typed Envelope & Taint-Gate Defense
Class 1: Multi-Step Data Exfiltration (EXFILTRATION)	Step 1: Read untrusted external document/issue → Step 2: Extract worker/workspace records into “diagnostic JSON” → Step 3: Invoke share_records or webhook tool.	Agent treats schema completion as a benign formatting task (100% compliance on share_records under framed_baseline).	Recognition + Provenance Taint Gate: Pre-turn recognition gate + untrusted-read taint tracking reduces share_records compliance from 100% → 0% .
Class 2: Confused Deputy Tool Execution (CONFUSED_DEPUTY)	Untrusted record embeds a pseudo-system remediation block instructing the agent to modify permissions, overwrite config files, or approve a flagged transaction.	Agent confuses passive data bytes with executable control instructions, invoking write/shell tools on behalf of the attacker.	Strict Data-vs-Control Separation: Passive files and retrieved records carry zero execution authority; tool calls require a pre-declared EffectAuthority receipt.

3. UNIFIED IMPLICATIONS FOR FRONTIER EVALUATION & AGENT SECURITY

Why Prompt-Only Guardrails Fail

Both Jurisdictional Confusion and Format-as-Affordance exploit the same transformer limitation: **attention is captured by local syntactic legitimacy** (well formed

The Recognition-First Verification Fix

Forcing a **1-sentence structured recognition check** (naming who pays, whether travel documents are held, and whether worker_records are shared) before generation